VALUE IN HEALTH 14 (2011) 1078–1084

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

CLINICAL OUTCOMES ASSESSMENT

Some Cautions on the Use of Instrumental Variables Estimators in Outcomes Research: How Bias in Instrumental Variables Estimators Is Affected by Instrument Strength, Instrument Contamination, and Sample Size

William H. Crown, PhD[1,], Henry J. Henk, PhD[2], David J. Vanness, PhD[3]*

[1]*OptumInsight Life Sciences, Waltham, MA, USA;* [2]*OptumInsight Life Sciences, Eden Prairie, MN, USA;* [3]*University of Wisconsin, Madison, WI, USA*

ABSTRACT

Objectives: To examine the performance of instrumental variables (IV) and ordinary least squares (OLS) regression under a range of conditions likely to be encountered in empirical research. **Methods:** A series of simulation analyses are carried out to compare estimation error between OLS and IV when the independent variable of interest is endogenous. The simulations account for a range of situations that may be encountered by researchers in actual practice—varying degrees of endogeneity, instrument strength, instrument contamination, and sample size. The intent of this article is to provide researchers with more intuition with respect to how important these factors are from an empirical standpoint. **Results:** Notably, the simulations indicate a greater potential for inferential error when using IV than OLS in all but the most ideal circumstances. **Conclusions:** Researchers should be cautious when using IV methods. These methods are valuable in testing for the presence of endogeneity but only under the most ideal circumstances are they likely to produce estimates with less estimation error than OLS.

Keywords: bias, endogeneity, instrumental variables.

Introduction

There is considerable interest in the potential application of instrumental variable (IV) techniques in the field of outcomes research. In large part, this can be traced back to the increased demand for real-world evidence regarding treatment effectiveness, safety, and value on the part of payer organizations and regulatory authorities. For example, the comparative effectiveness research component of health reform in the United States is specifically focused on understanding comparative treatment effectiveness and safety in actual clinical practice.

In any real-world analysis, there are a variety of conditions that can lead to biased estimates. In the vernacular of econometrics, these can all be classified under the general concept of endogeneity. In the simplest terms, endogeneity is defined as a nonzero correlation between an explanatory variable x_j and the disturbance term u of a regression equation. Such a correlation can be generated by a wide variety of sources, including omitted variables, measurement error, incorrect functional form, simultaneity, sample selection bias, and various combinations of these problems (note that endogeneity is similar to the concept of confounding in the field of epidemiology [1]). Endogeneity is a source of concern because, by definition, its presence means that parameter estimates from ordinary least squares (OLS) will be biased and inconsistent (see for example [2]). Moreover, the sources that generate it nearly always exist in studies involving observational data. To make matters worse, the researcher never knows how big the endogeneity problem is in any particular study because the disturbance term u is unobserved and, as a consequence, so is the extent of the correlation between the endogenous variable x_j and u.

Given its importance, it is not surprising that the topic of endogeneity has long been a central topic in the econometrics literature. IV approaches for addressing the problem of endogeneity date to the 1920s—although the identity of the inventor remains in doubt and will probably never be established for certain [3]. Early applications of IV focused on the estimation of supply and demand curves in agricultural markets. Economists typically observed only market clearing prices and quantities. Consequently, they needed statistical methods that would enable them to estimate the separate demand and supply relationships underlying markets. The IV approach relies on finding at least one variable that is correlated with the endogenous variable but uncorrelated with the outcome. In the case of estimating supply and demand for agricultural products, this meant coming up with variables that were related to demand but not supply (e.g., income), as well as those that were related to supply but not demand (e.g., weather conditions). With more than nine decades to accumulate, the theoretical and applied literature on IV estimation is vast [4].

In outcomes research applications, IV estimation methods are most often used to help address problems of sample selection bias. Sample selection bias arises when there are unmeasured variables that influence both treatment selection and outcomes. Failure to

* *Address correspondence to:* William Crown, OptumInsight Life Sciences, 950 Winter Street, Suite 3800, Waltham, MA, USA.
E-mail: bill.crown@innovus.com.

1098-3015/$36.00 – see front matter

doi:10.1016/j.jval.2011.06.009

account for these unmeasured variables results in biased estimates of treatment effects because the parameter estimate of the treatment variable also reflects the effects of the missing variables on outcomes. Conceptually, the objective of the IV approach is to estimate a model of treatment selection as a function of a set of independent, nonendogenous variables. At least one of these variables needs to be correlated with treatment selection but uncorrelated with the outcome variable. Because treatment is modeled as a function of nonendogenous variables, the predicted values of treatment from the model are, by construction, not correlated with the residuals of the outcome equation. Hence, by substituting the predicted values of the treatment variable for the original values when estimating the treatment effect in the outcome model the endogeneity problem has been solved and it is possible to obtain unbiased estimates of the treatment effect. However, the success of the IV approach hinges critically on the properties of the instrumental variables identified as being correlated with treatment selection but uncorrelated with the outcome variable.

One of the seminal papers on the use of IV methods in outcomes research is by McClellan et al. [5]. In that study the authors created a variable that indicated if the nearest hospital treated acute myocardial infarction admissions with cardiac catheterization. This assumed that distance was related to the treatment setting where patients received care but was unrelated to outcomes itself. Numerous other articles have subsequently used distance as an IV in outcomes studies [6–9]. Aside from distance, other IVs have included differential physician payment amounts associated with different treatments [6], relative time on market as an instrument for diffusion of knowledge to physicians about alternative treatments [10], out-of-pocket copayments for alternative pharmaceutical products [11], and many others. In each of these studies, there was reason to expect that the results of standard OLS regression estimates would be biased by endogeneity. In each instance, the authors were able to identify an IV that was correlated with the explanatory variable of interest and assumed to be uncorrelated with the disturbance term of the outcome equation. But the fact that the IV estimates may have differed from the OLS results doesn't necessarily mean that the IV estimates were less biased. As we demonstrate empirically in this article using simulation methods, the relative unbiasedness and estimation error of IV and OLS hinges critically on a number of factors, including the strength of the instrument and the degree to which the instrument still retains some correlation with the error term of the outcome equation. Larger sample sizes make the distinction between the properties of IV and OLS under various conditions more apparent but do not fundamentally change conclusions about the relative merits of the two estimators under a given set of conditions.

Despite the appeal of IV methods for addressing the many variants of endogeneity that commonly arise in the analysis of observational data, researchers have raised concerns over the performance of IV and parametric sample selection bias models—noting, in particular, the practical problems often encountered in identifying good instruments, sensitivity of instruments to assumptions about functional form [12], and general lack of understanding of the statistical properties of IV in nonlinear models (except for the handful of cases where these properties have been established). It is remarkably difficult to come up with strong instruments (i.e., variables that are highly correlated with the endogenous variable) that are uncorrelated with the disturbance term. As a result, instruments tend to be either weakly correlated with the variable for which they are intended to serve as an instrument, correlated with the disturbance term, or both.

Staiger and Stock [13] note that empirical evidence on the strength of instruments is sparse. In their review of 18 articles published in the *American Economic Review* between 1988 and 1992 using two-stage least squares, none reported first stage *F* statistics or partial R^2s measuring the strength of identification of the instruments. Although not reported in the original studies, Ebbes [14] found considerable variability in the strength of instruments used by researchers in the returns to education literature [15,16]. More recently, researchers have been more likely to report statistics related to instrument strength and contamination. For example, several articles in the health services research area have reported on the strength of the instruments used [5,9,10].

Although it is impossible to draw any systematic conclusions from this handful of examples, it is safe to assume that 1) IV methods will often generate different conclusions than standard regression models that do not attempt to control for endogeneity; 2) the strength of instruments used in various studies varies widely; and 3) instruments are often correlated with the disturbance term of the outcome equation. In fact, we argue that in an effort to minimize the correlation of the instrument with the disturbance term, researchers will have a tendency to identify weak instruments.

In general, the more highly an instrument is correlated with the endogenous variable, the more likely that the instrument is also correlated with the unobservable u [17]. Consider a simple linear model where the outcome variable y is a linear function of a treatment variable x, a parameter β and an unobservable error term e_1:

$$y = x\beta + e_1, \quad \text{where } E(e_1) = 0. \tag{1}$$

Suppose also that the treatment variable x is, in turn, a function of another variable z, a parameter $\Gamma \neq 0$ and an unobservable e_2 (assumed to be uncorrelated with the error term of the outcome equation e_1:

$$x = z\Gamma + e_2 \quad \text{where } E(z\, e_1) = 0,\ E(e_2) = E(z\, e_2) = 0 \tag{2}$$

Suppose also that $\rho(e_1,e_2) \neq 0$. The presence of e_2 in the determination of x gives rise to the endogeneity problem through its correlation with e_1. Assuming it is measurable, the variable

$$v = z\Gamma + \alpha e_2 \tag{3}$$

arises as a natural choice of instrumental variable for x. The "strength" of the instrument v may be expressed by its correlation with x, $\rho(v,x)$. Comparing equations (2) and (3) the strongest possible instrument for x is x itself (i.e., $\rho(x,v) = 1$ when $\alpha = 1$). However, because $\rho(e_1,e_2) \neq 0$, the instrument is only "uncontaminated" (i.e., $\rho(v,e_1) = 0$) when $\alpha = 0$, and furthermore, the level of contamination is increasing in α. In other words, the two desirable characteristics of a "good" instrument (strong covariance with the endogenous explanatory variable and no correlation with the unobservable) seem to be inherently in tension. The reason that strong instruments and clean instruments are not theoretically incompatible is that, in addition to its covariance with e_1, e_2 contains a random component μ that is uncorrelated with e_1.

Based on the above, it is reasonable to expect that researchers would gravitate toward the use of weak instruments to reduce the chance of using an instrument that is itself endogenous. Several studies, however, have shown that weak instruments may lead not only to larger standard errors in treatment estimates but may, in fact, lead to estimates that have larger bias than OLS [13,18–20]. In particular, Bound et al. [18] show that the incremental bias of IV versus OLS is inversely proportional to strength of the instrument and the number of excluded variables. In this article we investigate if the Bound et al. [18] results give rise to practical concerns for empirical researchers by simulating how bias in an IV estimator is related to the strength of the correlation between the instrument and the variable that it is intended to replace. We also examine how bias in the IV estimator is related to the strength of the correlation between the instrument and the observed residuals (the contamination of the instrument). Finally, rather than considering

the behavior of these estimators in the limit, we examine how bias changes in relation to sample size for a range of study sizes more likely to be encountered in practice.

Methods

In the next section we present the results of a simulation study to demonstrate estimation error caused from correlation between the disturbance term and the IV when sample selection models are employed to estimate regression parameters in the presence of an endogenous explanatory variable. In what follows, we assume the purpose of the analysis is to estimate the average treatment effect. There has been considerable discussion in the literature about what is meant by treatment effects—particularly when there is heterogeneity in response to treatment (see for example [21]). For the purposes of the simulations presented in this article, average treatment effects are assumed not to be heterogeneous. By average treatment effects we mean the difference in expected outcomes (shift in the constant term) when the treatment dummy T is 1 versus 0. T is a determined by a latent treatment selection model:

$x > 0 \quad T = 1$

else $\quad T = 0$

where $x = e_2$.

Consider a model of outcome, y, as a function of treatment, T and an unobservable e_1

$y = \beta_0 + \beta_1 T + \varepsilon \;\; where \;\; \varepsilon = e_1.$

We also construct an instrumental variable, z, which, ideally, would be correlated with T, but not with e_1.

We generate a trivariate normal variance covariance distribution (e_1,e_2,z) where e_1 are the residuals from the outcome equation, e_2 are the residuals from the treatment selection equation, and z is the instrumental variable. For illustrative purposes, we set the covariance of the residuals in the outcome and treatment selection equations at. This strikes us intuitively as a substantial level of endogeneity for the purposes of the simulations. Next, we set the correlation of the instrument with treatment selection to be either 0.5 (a "strong" instrument) or 0.25 (a "weak" instrument). Note that, for the purposes of the simulations, it is necessary only to define instruments that are strong and weak in a relative sense. Our strong instrument is defined to have twice the correlation with treatment than the weak instrument. In practice, defining strong and weak instruments is generally based upon the significance of the parameter estimates in the first stage equation relating the instruments to treatment. However, aside from statistical significance, some instruments may be much stronger than others. We wish to simulate the effect of using relatively stronger instruments versus weaker instruments to address a given level of endogeneity (Baiocchi et al. [22] discuss the tradeoff between instrument strength and sample size. They find that using observations at the two extremes of the excess travel time distribution by discarding the middle observations considerably strengthens excess travel time as an instrument. This effect outweighs the loss of sample size). We then allow the correlation between the instrument and unobservable to vary across a range of values. This allows us to assess the effects of remaining endogeneity if the instrument is not completely independent of the residuals. Finally, outcomes y are generated, given (e_1,e_2,z), with true values $\beta_0 = 6$ of and $\beta_1 = 1$.

We model the covariances among the errors as trivariate normal because this conforms to the majority of the literature on IV and sample selection models. The literature on sample selection models generally assumes an OLS model with a normal distribution for the residuals in the outcome equation (1) and a cumulative normal probability distribution for the error distribution in equation (2).

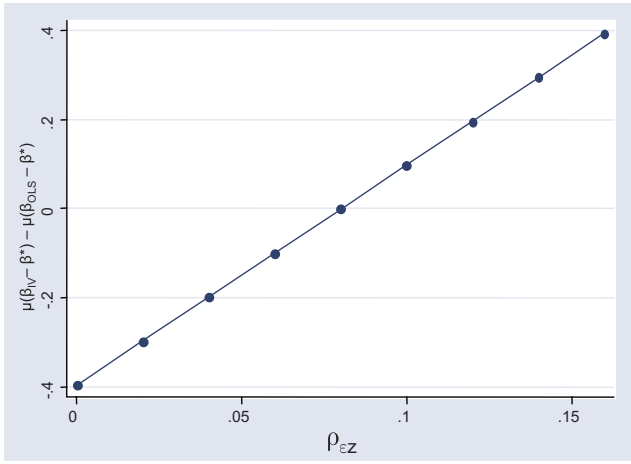

Fig. 1 – Mean estimation error as a function of $\rho_{\epsilon z}$.

The number of replicate studies was chosen [23] to be large enough to illustrate the effects of alternative assumptions about strength of endogeneity, strength of the instrument, correlation of the instrument with the residuals of the outcome equation, and sample size. Because treatment is a binomial variable we can use standard power tables to estimate the required sample size to detect alternative effect sizes in the treatment estimates. A sample size of 1000 is sufficient to detect effect sizes of 20% or larger at a 99% confidence level with a 97% power. Two previous important simulation articles in the field chose the number of replicate studies to be 500 [24] and 1000 [25].

The analyses were programmed using Stata 8.2 (200x, Stata Corp, College Station, TX).

Results

The strong instrument case

In the strong instrument case, we find that differences in the mean estimation error (i.e., bias) $\mu(\beta_{IV} - \beta^*) - \mu(\beta_{OLS} - \beta^*)$ increase linearly as the contamination of the instrument (i.e., $\rho_{\varepsilon z}$) increases, with an intercept approximately equal to $-\rho_{\varepsilon T}.(\sigma_\varepsilon/\sigma_T) = -0.4$ and slope approximately equal to $(1/\rho_{Tz}).(\sigma_\varepsilon/\sigma_T) = 5$ consistent with the results of Bound et al. [18]. Note that the mean estimation error of the IV estimator equals or exceeds the mean estimation error of the OLS estimator (i.e., is > 0) approximately when $\rho_{\varepsilon z} \geq \rho_{Tz} \cdot \rho_{\varepsilon T} = 0.08$ (see Fig. 1).

It is important to note that this result with respect to bias is only part of the story. This is because neither asymptotic nor average performance of the IV estimator relative to OLS necessarily reflects performance for a particular study with a finite sample size. To illustrate this fact, we plot the distribution of estimation errors from 1000 replicate studies, each with a sample size of n = 2000, for three levels of instrument contamination: $\rho_{\varepsilon z} = 0.0$, 0.08 and 0.16 (see Fig. 2).

With this relatively large sample size, the sampling distributions of the OLS and IV estimators have small standard errors and the distributions do not overlap. IV outperforms OLS (i.e., has lower estimation error) in 1000 out of 1000 replicate studies with a perfectly clean instrument ($\rho_{\varepsilon z} = 0$). Conversely, OLS outperforms IV in 1000 out of 1000 replicate studies with a very contaminated instrument ($\rho_{\varepsilon z} = 0.16$). Furthermore, even though the IV estimator

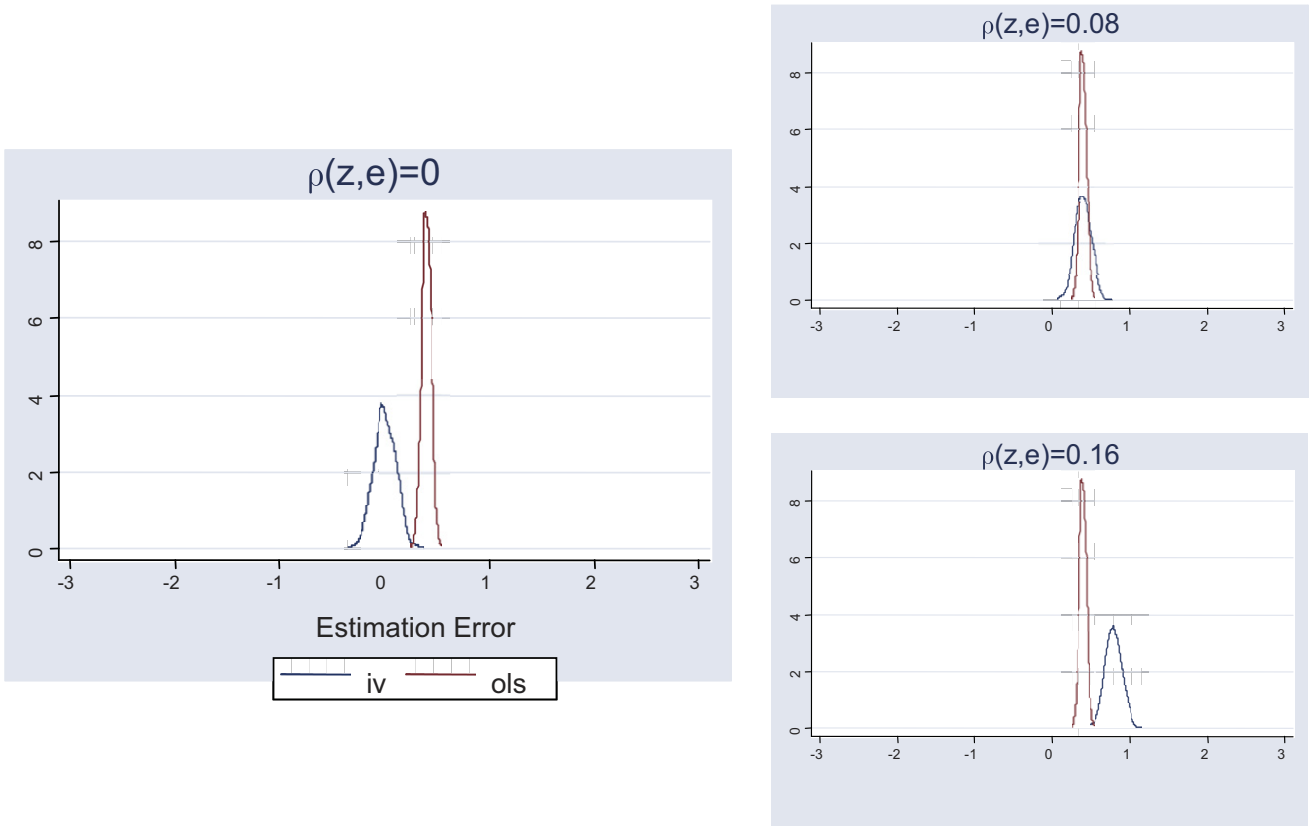

Fig. 2 – Distribution of estimation error (large sample case).

was unbiased, the standard deviation of the empirical IV sampling distributions was nearly two and a half times that for the OLS estimator (0.106 vs. 0.043, 0.104 vs. 0.043, and 0.106 vs. 0.043, for $\rho_{\varepsilon z}$= 0.0, 0.08, and 0.16, respectively). Perhaps most importantly, we note that the sign of the OLS estimation error is consistently positive, while the IV estimation error is sometimes positive and

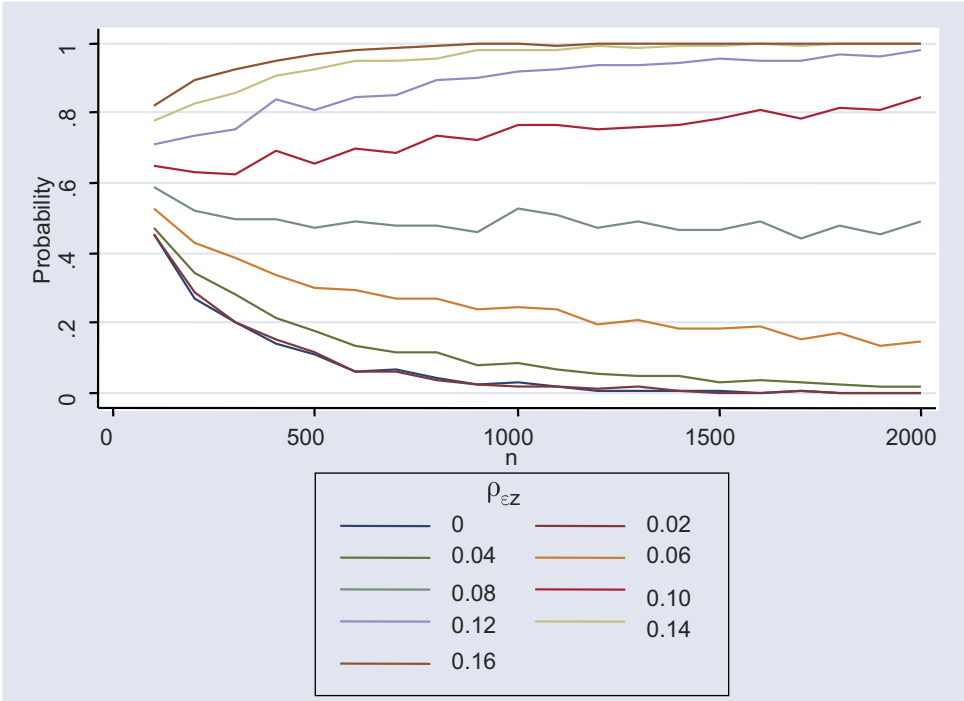

Fig. 3 – Proportion of replications where $|(\beta IV - \beta^*)| - |(\beta OLS - \beta^*)| > 0$.

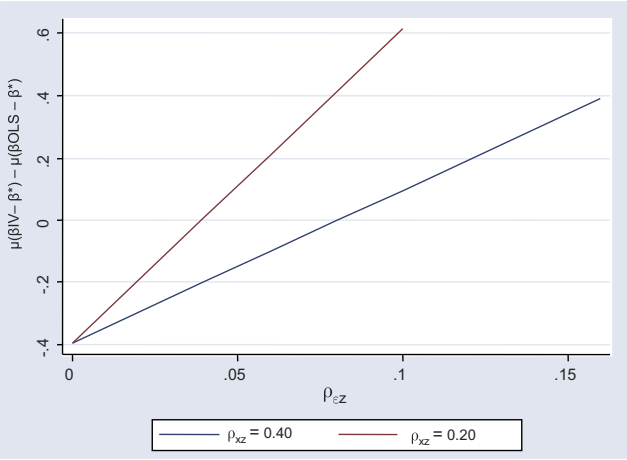

Fig. 4 – Mean estimation error as a function of $\rho_{\epsilon z}$.

sometimes negative. As a result, it is not possible to have a expectation about the direction of bias with IV when the IV is not completely exogenous.

The convergence of IV performance relative to OLS at various levels of instrument contamination can be clearly visualized by plotting the probability that IV is outperformed by OLS (i.e., the percent of the 1000 replicate studies in which $|(\beta_{IV} - \beta^*)| - |(\beta_{OLS} - \beta^*)| > 0$ as the sample size of each replicate increases (Fig. 3). (Technically, these are estimated relative frequencies rather than probabilities because they are generated by replicate samples rather than actual population). The asymptotic results indicate that all lines below 0.08 will eventually approach zero, whereas all lines above 0.08 will eventually approach one.

Asymptotic results do not necessarily indicate better performance in any particular study until sample size is relatively large—even for fairly uncontaminated instruments. For example, with a relatively uncontaminated instrument ($\rho_{\varepsilon z} = 0.06$), the probability that IV is outperformed by OLS in any given study remains as high as 25% for n = 1000 and is still over 15% for n = 2000.

The weak instrument case

We next consider the case of a weaker instrument in which the correlation between z and the endogenous latent selection variable, x, equals the correlation between x and the unobservable, ε (i.e., $\rho_{xz} = 0.25$ and $\rho_{\varepsilon x} = 0.25$). We note that even this weaker instrument is fairly strong relative to the magnitude of the endogeneity problem, say, in comparison to the classic study by Angrist and Krueger [26] in which birth quarter served as an instrument for educational attainment in an estimate of the effect of education on weekly earnings.

We again find a linear relationship between the difference in mean estimation error $\mu(\beta_{IV} - \beta^*) - \mu(\beta_{OLS} - \beta^*)$ and the contamination of the instrument (i.e., $\rho_{\varepsilon z}$), with IV estimation bias exceeding OLS estimation bias approximately when $\rho_{\varepsilon z} \geq \rho_{Tz} \cdot \rho_{\varepsilon T} = 0.04$ (see Fig. 4). In comparison to Figure 1, the relationship displayed in Figure 4 has a slope twice as steep, due to the halving of ρ_{Tz}, whereas the intercept remains the same.

We next demonstrate that finite sample performance of the IV estimator relative to OLS can deviate even further from asymptotic results when the instrument is relatively weak. We again plot the distribution of estimation error from 1000 replicate studies, each with a sample size of n = 2000, for three levels of instrument contamination: $\rho_{\varepsilon z} = 0.0$, 0.04 and 0.08 (see Fig. 5), noting that the

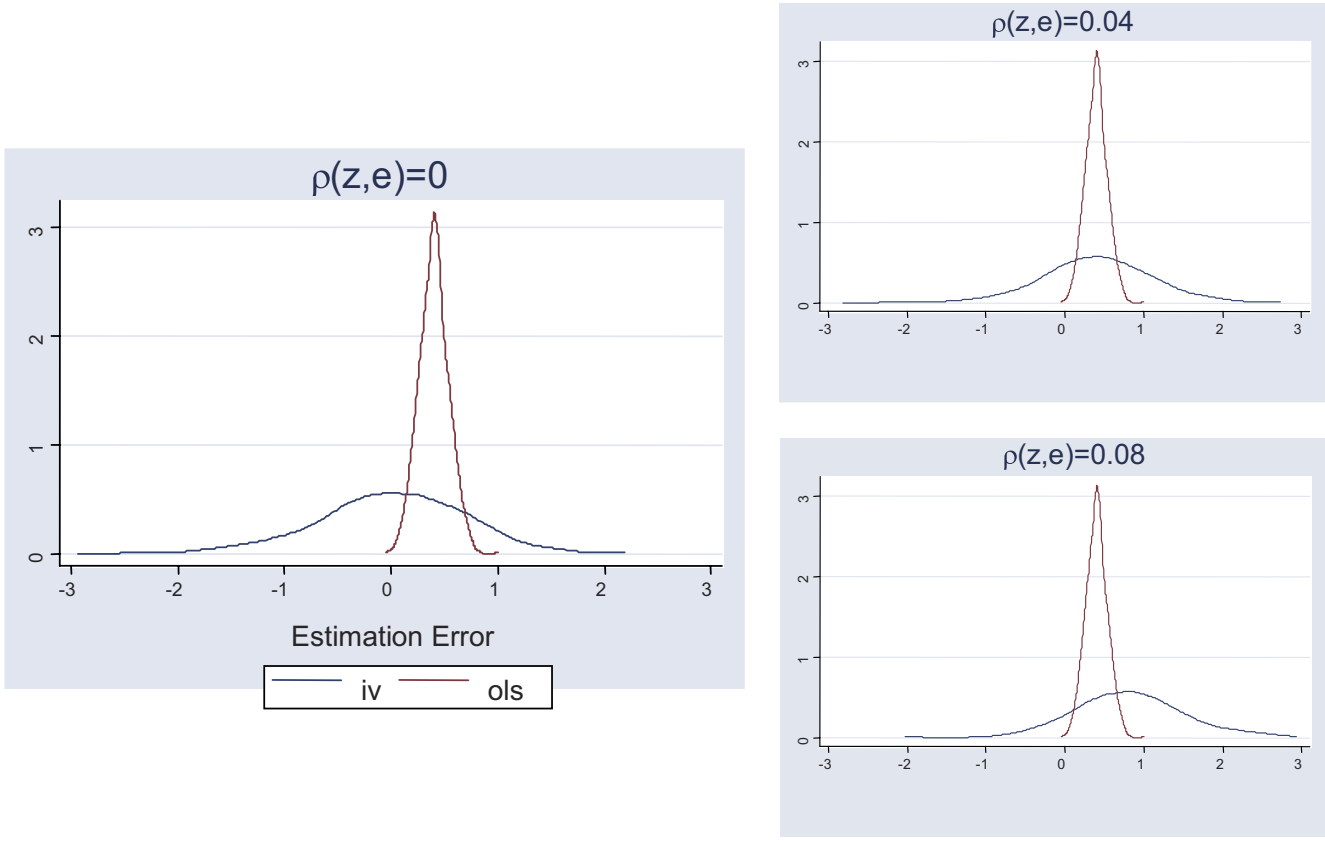

Fig. 5 – Distribution of estimation error (large sample case).

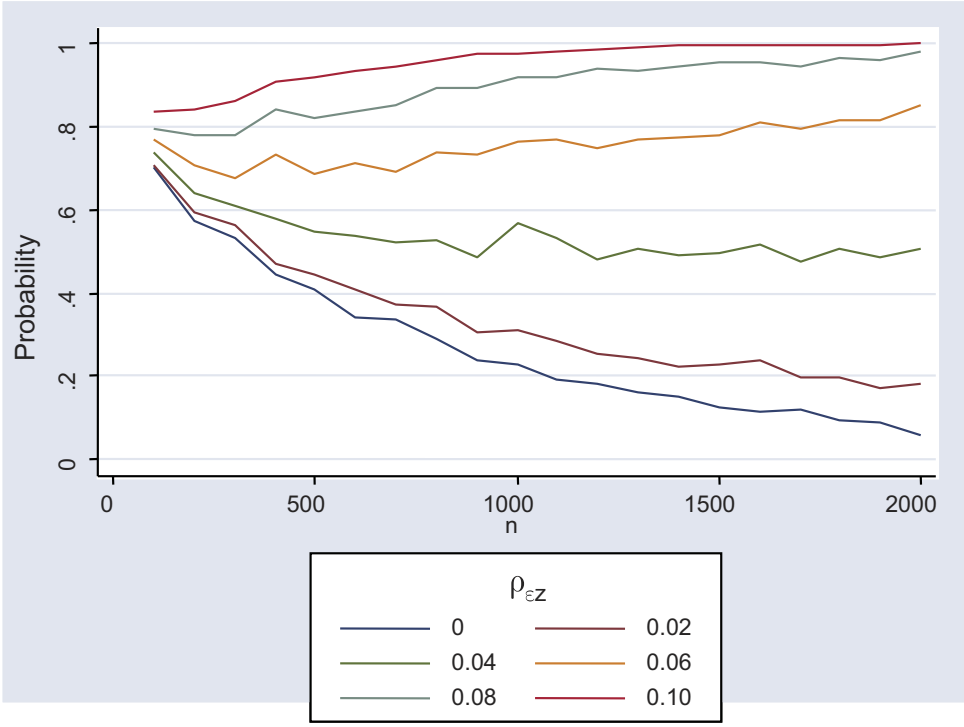

Fig. 6 – Probability $|(\beta_{IV} - \beta^*)| - |(\beta_{OLS} - \beta^*)| > 0$.

levels of instrument contamination chosen are now exactly half the levels in the strong instrument case.

With a sample size of n = 2000, we find simulated finite sample performance that is not as close as when the instrument was strong (see Fig. 5).

The distributions of the OLS and IV estimation errors are noticeably more dispersed than in the strong instrument case. OLS, however, still outperforms IV (i.e., has lower estimation error) in 59 out of 1000 replicate studies, even with a perfectly clean instrument ($\rho_{\varepsilon z} = 0$). With a sample size of 2000 for each replicate, the ratio of the standard deviation of IV estimation errors to the standard deviation of OLS estimation errors has settled to approximately five to one (0.215 vs. 0.043, 0.212 vs. 0.043, and 0.219 vs. 0.043, for $\rho_{\varepsilon z} = 0.0$, 0.04, and 0.08, respectively). In other words, at these sample sizes, the ratio of the standard deviation of IV estimation errors to the standard deviation of OLS estimation errors approximately doubles when the strength of the instrument is cut in half.

This result leads us to expect slower convergence of the estimators' relative performance to their asymptotic results. Again, we plot the probability that IV is outperformed by OLS (i.e., the percent of the 1000 replicate studies in which $|(\beta_{IV} - \beta^*)| - |(\beta_{OLS} - \beta^*)| > 0$) as the sample size of each replicate increases (Fig. 6).

That the bias of IV is lower than that of OLS for instruments with contamination levels $0 \leq \rho_{\varepsilon z} < 0.04$ says little about relative performance of the methods in a particular study until sample size is quite large. Even with a perfectly uncontaminated instrument ($\rho_{\varepsilon z} = 0$), the probability that IV is outperformed by OLS in any given study remains over 20% for n as large as 1000. For a very slightly contaminated instrument ($\rho_{\varepsilon z} = 0.02$), OLS still outperformed IV in nearly 20% of the studies, even with n = 2000.

Conclusions

There is a large econometrics literature using simulation methods to evaluate the empirical properties of alternative estimators. However, to our knowledge there are no studies comparing OLS and IV under alternative assumptions regarding instrument strength, sample size, degree of the endogeneity problem, and correlation of the instruments with the residuals in the outcome equation. The simulation results reported in this article are intended to provide empirical estimates of the magnitude of estimation error under various conditions using the theoretical framework of Bound et al. [18] as the guiding structure. The intent of this article was to provide researchers with more intuition around how important these issues are from an empirical standpoint. In particular, it is rare to see discussion of the potential effects of instrument contamination in the applied health econometrics literature. Rather common, on the other hand, are heroic attempts to "find an instrument" whenever the slightest possibility of endogeneity arises.

Although the appeal of IV as a method for addressing endogeniety issues is undeniable, it is important to understand that the use of IV can do more harm than good. In fact, the simulations indicate a greater potential for inferential error when using IV than OLS in all but the most ideal circumstances. We have shown that even the sobering asymptotic results of Bound et al. [18], demonstrating the maximally acceptable level of contamination for an instrument relative to both the instrument's strength and the seriousness of the endogeneity problem, are perhaps not sobering enough. In actual empirical work, finite sample sizes affect the variance of the distribution of estimation errors and this is compounded when the instrument is weak. For the size of samples used in most studies, the probability that IV is outperformed by OLS is substantial, even when the asymptotic results indicate bias to be lower for IV.

Unfortunately, just as we can never know the true magnitude of the endogeneity problem to begin with, we can never know exactly how contaminated our instrument really is (both quantities depend on the unobservable residuals of the outcome equation). We do, however, have the ability to measure the correlation of our instrument with observed treatment selection. And making some assumptions about the properties of

the empirical residuals relative to the unobserved ε we can certainly test for endogeneity, contamination of the instrument, and instrument strength. Brookhart et al. [27] provide an excellent set of practical recommendations for conducting and reporting IV analyses.

Perhaps less widely recognized, it is possible to infer the likely direction of the bias in OLS estimates due to the presence of correlated unobservables (that we can often name but cannot measure). Based upon the simulation results, we note that the mean of the estimated sampling distribution for OLS was consistently greater than the known parameter value being estimated. As a result, the direction of bias from OLS is consistently positive and it can be assumed that the bias from OLS is positive. On the other hand, the mean of the sampling distribution for IV was sometimes greater than the known parameter value and sometimes less. As a result, the direction of bias with IV when the instrument is correlated with the residuals cannot be determined a priori.

This seems like a fairly important piece of information to give up.

We urge caution in using IV methods in treatment effects regression at even the merest suggestion of endogeneity. If it is possible to identify and construct a plausible instrument, there may well be value in using IV as a test for the presence of endogeneity. Researchers should recognize that they will need to find a strong instrument and it will need to have a very low correlation with the empirical residuals. And they should make sure that they have a large sample.

REFERENCES

[1] Zohoori N, Savitz D. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. Ann Epidemiol 1997;7:251–7.
[2] Wooldridge J. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press, 2002.
[3] Stock JH Trebbi F. Who invented instrumental variables regression? J Econ Perspect 2003;17:177–94.
[4] Murray M. Avoiding invalid instruments and coping with weak instruments. J Econ Perspect 2007;20:111–32.
[5] McClellan M, McNeil B, Newhouse. J. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? JAMA 1994;272:859–66.
[6] Hadley J, Polsky D, Mandelblatt JS, et al. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a Medicare population. Health Econ 2003;12:171–86.
[7] McConell K, Newgard C, Mullins R, et al. Mortality benefit of transfer to level I versus level II trauma centers for head-injured patients. Health Serv Res 2005;40:435–57.
[8] Brooks J, Irwin C, Hunsicker L, et al. Effect of dialysis center profit-status on patient survival: a comparison of risk-adjustment and instrumental variable approaches. Health Serv Res 2006;41:2267–89.
[9] Brooks J, Chrischilles E. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. Med Care 2007;45(Suppl. 2):S123–30.
[10] Crown W, Hylan T, Meneades L. Antidepressant selection and use and healthcare expenditures. Pharmacoeconomics 1998;13:435–48.
[11] Cole J, Norman H, Weatherby L, et al. Drug copayment and adherence in chronic heart failure: effect on cost and outcomes. Pharmacotherapy 2006;26:1157–64.
[12] Goldberger A. Abnormal Selection Bias. In: Karlin S, Amemiya T, Goodman L, eds. Studies in Econometrics, Time Series, and Multivariate Statistics. New York: Academic Press, 1983.
[13] Staiger D, Stock JH. Instrumental variables regression with weak instruments. Econometrica 1997;65:557–86.
[14] Ebbes P. Latent Instrumental Variables—A New Approach to Solve for Endogeneity [dissertation]. Groningen, The Netherlands: Rijksuniversiteit Gronigen, 2004.
[15] Mroz TA. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica 1987;55:765–99.
[16] Card D. Using geographical variation in college proximity to estimate the return to schooling. In: Christofides LN, Grant E, Swidinsky R, eds. Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp. Toronto: University of Toronto Press, 1995.
[17] Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 2002; 20:518–29.
[18] Bound J, Jaeger DA, et al. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. JASA 1995;90:443–50.
[19] Hahn J, Hausman J. A new specification test for the validity of instrumental variables. Econometrica 2002;70:163–89.
[20] Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variables regression. J Econ 2003;114:29–72.
[21] Basu A, Heckman J, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ 2007; 16:1133–57.
[22] Baiocchi M, Small D, Lorch S, Rosenbaum P. Building a stronger instrument in an observational study of perinatal care for premature infants. J Am Stat Assoc. 2010;105:1285–96.
[23] Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
[24] Basu A, Manning W, Mullahy J. Comparing alternative models: log vs cox proportional hazard? Health Econ 2004;13:749–65.
[25] Manning W, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ 2001;20:461–94.
[26] Angrist J, Krueger A. Does compulsory school attendance affect schooling and earnings. Quarter J Econ 1991;106:979–1014.
[27] Brookhart MA, Rassen JA, Schneeweis S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010:19;537–54.